AD-A250 978

DTIC
ELECTE
MAY 28 1992
S C D

THE BASIC BAYESIAN BLUNDER*

by

Henry E. Kyburg, Jr.

Abstract

"Direct Inference" was distinguished from "Inverse Inference" early in the development of mathematical statistics. Direct inference was the form of uncertain inference that took as premise a distribution in a population, and yielded a (probable) conclusion about the composition of a sample from that population. Inverse inference was to take as a premise the composition of a sample, and to yield as a conclusion a (probable) conclusion about a distribution in a population. Direct inference seemed uproblematic. But inverse inference seemed to be needed to obtain the general premises needed for direct inference. Inverse inference proper is based on Bayesian principles. This paper argues that these principles are inconsistent with direct inference. It is concluded that we should hold fast to direct inference, and accept Bayesian procedures only when they can be put into the framework of direct inference.

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

92-13700

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1989	Unknown

4. TITLE AND SUBTITLE

The Basic Bayesian Blunder

5. FUNDING NUMBERS

DAAB10-86-C-0567

6. AUTHOR(S)

Henry E. Kyburg, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Rochester
Departmenet of Philosophy
Rochester, NY 14627

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army CECOM Signals Warfare Directorate
Vint Hill Farms Station
Warrenton, VA 22186-5100

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

92-TRF-0025

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Statement A; Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

"Direct Inference" was distinguished from "Inverse Inference" early in the development of mathematical statistics. Direct Inference was the form of uncertain inference that took as premise a distribtion in a population, and yiedled a (probable) conclusion about the composition of a sample from the population. Inverse inference was to take as a premise the composition of a sample, and yield as a conclusion a (probable) conclusion about a distribution in a population. Direct inference seemed uproblematic. But inverse inference seemed to be needed to obtain the general premised needed for direct inference. Inverse inference proper is based on Bayesian principles. This paper argues that these principles are inconsistent with direct inference. It is concluded that we should hold fast to direct inference, and accept Bayesian procedures only when they can be put into the frame work of direct inference.

14. SUBJECT TERMS

Artificial INtelligence, Data Fusion, Philosophy, Bayesian

15. NUMBER OF PAGES

21

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to **stay within the lines to meet optical scanning requirements.**

Block 1. Agency Use Only (Leave blank).

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. (If known)

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - DOD - Leave blank.
DOE - DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - NASA - Leave blank.
NTIS - NTIS - Leave blank.

Block 13. Abstract. Include a brief (Maximum 200 words) factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code (NTIS only).

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

I

If any distinction in the realm of statistics or inductive logic deserves to be called "classical", the distinction between direct and inverse inference does. In philosophy, it is the classic distinction between inductive and deductive argument. Inferring that Socrates the man is mortal from the premise that all men are mortal, is an instance of direct inference. The corresponding inverse inference is that which proceeds from premises, Socrates the man is mortal, Plato the man is mortal, ... Churchill the man is mortal, to the general conclusion that all men are mortal. Inverse inference proceeds from the particular to the general, direct inference from the general to the particular. Inverse inference is characterised by inductive logic; direct inference by deductive logic.

In statistics the distinction is even more straight-forward: it is the distinction between inferences that take knowledge of a distribution in a population as a premise, and infer the probable character of a particular sample -- this is direct inference -- and inferences that take knowledge of a sample as a premise, and infer the probable character of the population from which the sample comes -- this is inverse inference.

Inverse inference is characterized by what in artificial intelligence is called non-monotonicity. This means that, in contrast to deductive inference, an increase in the premises may undermine a conclusion already reached. This was recognized explicitly by R. A. Fisher in 1936. He writes, "There is one peculiarity of uncertain inference which often presents a difficulty to mathematicians trained only in the technique of rigorous deductive argument, namely that our conclusions are arbitrary, and therefore invalid, unless all the data, exhaustively, are taken into account. In rigorous deductive reasoning we may make any selection from the

data, and any certain conclusions which may be deduced from this selection will be valid, whatever additional data we have at our disposal."[1]

In general, whether in logic or in statistics, direct inference has been regarded as relatively unproblematic. Basic, first order, deductive logic is almost universally accepted as being all right as far as it goes, though there are some people who think it does not go far enough. (It does not capture modal arguments, for example.)

In a similar way, early on in the history of probability theory agreement was achieved concerning the inferences that were warranted whose premises concerned general distributions, and whose conclusions concerned samples. Given as a premise that heads among coin tosses are distributed binomially, with a parameter of 1/2, we all easily calculate that the probability of four heads in succession is 1/16. There is no uncertainty in the argument here. But there are also direct inferences that embody uncertainty: We infer -- or, as Neyman[2] for example might prefer to put it, we <u>behave</u> as if -- the next ten tosses of this coin will not yield ten heads.

Given as premises however, the distribution of heads in a quite large sample of tosses, subjected to whatever constraints concerning randomness you wish, there is inevitably controversy concerning what conclusion is warranted and to what degree.

Note that in statistics both the direct and inverse inference may be non-monotonic: to augment the premises may undermine direct uncertain inference as well as the inverse inference. To learn that not only are the ten tosses we are concerned with the <u>next</u> ten tosses, but that nine of them have already yielded heads undermines our inference from the general distribution to the conclusion that we won't get ten heads in a row.

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

Similarly, given the results of n tosses, the inference (whatever it may be) to a general distribution of heads among tosses will be undermined (rendered epistemically irrelevant) by knowledge of the outcomes of an additional m tosses. This fact is of interest, particularly when it comes to constructing a logic that will reflect the realities of uncertain inference. But it is not essential to the distinction between direct and inverse inference. What is essential is that in direct inference the movement is from the population to an actual or hypothetical sample, while in inverse inference the movement is from the sample to some statement concerning the parent population.

II

What is special about inverse inference is not the use of Bayes' theorem. When Neyman writes[3] "...persons who would like to deal only with classical probabilities, having their counterparts in the really observable frequencies, are forced to look for a solution of the problem of estimation other than by means of the theorem of Bayes," we must understand him to be emphasizing the phrase "solution to the problem of estimation," since Bayes' theorem is, after all, a theorem.

What this means is that, as R. A. Fisher saw clearly, there are many situations in which Bayes' theorem is applicable that can easily be construed in terms of direct inference. In 1930[4] he notes that drawing from a super-population in which the parameter of interest (say θ) has a known distribution F_θ, and then getting a posterior distribution G_θ for θ is "...a perfectly direct argument...", though of course it uses Bayes' theorem. In the same way the famous example described by La Place, concerning n + 1 urns, each containing n black and white balls in each

possible combination involves Bayes' theorem, but makes no use of inverse inference proper. (The application of this model to sampling does.)

For inverse inference proper -- that is, inference whose uncertainty is not based on known frequencies -- Fisher has nothing but contempt[5]: "In fact, the argument runs somewhat as follows: a number of useful but uncertain judgments can be expressed with exactitude in terms of probability; our judgments respecting causes or hypotheses are uncertain, therefore our rational attitude towards them is expressible in terms of probability." Neyman's attitude is even less tolerant.

Fisher and Neyman were, of course, reacting against the use of the so-called "axiom" of Bayes that gave uniform priors, against La Place's principle of indifference, and the like. Since that time, however, inverse inference has become respectable again. It gained respectability by admitting what De Finetti called its "subjective sources" and claiming nevertheless to provide a rationale for inferences from a sample to a population, thus completing, in a sense, the theory of statistically uncertain inference. Direct inference governs the inference from the population to the sample; indirect inference, by means of Bayes' theorem, governs the inference from the sample to the population.

I claim that there is a serious blunder involved here -- not quite so obvious as the fallacy Fisher offers us, but a blunder nevertheless. It lies in the fact that direct inference and inverse inference cannot coexist happily. Historically, we were all confident and happy with the use of direct inference. A number of people had philosophical qualms about its application to specific objects and events: "the next toss," "the next sample of a thousand balls to be drawn," etc. Nobody had formulated careful rules of application for direct inference; but few people doubted that direct inference was in principle sound: if you know that a coin is fair,

you can infer that the probability that the next toss will yield heads is a half, that the probability is a sixteenth that the next sample of four consecutive tosses will have the structure HHTH, etc. Those who had serious qualms replaced talk of probability with talk of confidence, inductive behavior, rules, etc.

What we wanted, and didn't have, was a generally acceptable rationale for inverse inference. That is what Thomas Bayes sought, and what the modern philosophical Bayesian seeks. But inverse inference proper undermines direct inference. In order to have inverse inference, to "complete" our theory of uncertain inference, we find we must abandon direct inference in many, if not all, of its classical applications.

III

The most elementary example of this conflict can be seen in the case of a simple binomial distribution. If we know that a coin is fair and that its tosses are independent, we have no difficulty in calculating the probability of, say, ten heads on the next fifteen trials. This is our old, unproblematic, direct inference.

But where does this knowledge come from? Inverse inference, so the story goes. That is, we suppose that the way we got our binomial hypothesis was by looking at a lot of coin tosses. So let H be the hypothesis that the coin is fair. If we "get" H by inverse inference, that cannot mean that we assign it probability one: inverse inference via conditionalization can't raise a probability to one that doesn't start there. But if the probability of H is not one, then all our conventional direct inferences are undermined. In particular, we can no longer regard the tosses as independent, since every toss will change (by conditionalization) the probability we assign to

$\binom{15}{10}$

that the tosses are independent, but this just means that any
on of a specified sequence occurs just as often as that same
The dependence among tosses is epistemic, and depends on our
epistemic state with regard to H. We can no longer just say that the probability of ten heads among 15 trials is $\binom{10}{15}\left(\frac{1}{2}\right)^{15}$ -- for (a) the very first head will change the probability we assign to H itself, and (b) we must also take account of the alternative hypothesis not-H, and what it assigns to ten heads.

IV

It has been claimed that both R. A. Fisher's fiducial inference, in the case of estimating the mean for a normal distribution, and Neyman's method of confidence intervals for estimating the mean of a binomial distribution, require a "flat" or "uninformative" prior distribution for their validity, and therefore are merely special cases of Bayesian inverse inference. As Fisher and Neyman, respectively, have pointed out, this is untrue. Whatever be the mean μ of the normal population, the quantity $(x - \mu)/\sigma$ will be normally distributed with unit variance and mean 0. For confidence interval estimation, whatever be the binomial parameter p, the frequency with which a sample will fall in the confidence region will be at least as great as the confidence level.

The germ of falsehood -- or better, irrelevance -- in this observation lies in the fact that if we had some prior distribution that was not the flat or uninformative distribution, then the fiducial or confidence argument would not be valid. But one should distinguish between knowing that p is uniformly distributed between 0 and 1, and knowing that p has some (totally unknown) value in that interval. This is the classic -- but not always

helpful -- distinction between an unknown constant and a random quantity.[6] To undermine the fiducial or confidence-interval inference, we must have positive knowledge that the prior distribution is not the flat or uniformative prior. Approaching the question from the other side, no matter what prior distribution we feed into Bayes' theorem, we can worry that it requires (presupposes) some corresponding frequency or propensity. But of course no subjectivistic Bayesian would share these worries.

Now if the prior distribution is a frequency-like distribution in some super-population, then it is merely that a different direct inference is called for (as Fisher and Neyman both saw), and we aren't talking about inverse inference proper. Direct inference will do just as well. But if the prior distribution is a priori or subjective, as it must be for an inverse inference proper, then there is conflict between the inverse inference and the fiducial or confidence argument based on direct inference.

V

Inverse inference and direct inference will agree if the prior distribution provided by the inverse inference happens to be the uninformative prior. I have argued elsewhere that this fortunate situation is relatively rare. In a previous paper I used a procedure of de Finetti to show that it is very easy for general empirical hypotheses to achieve impressively high probabilities in the absence of any evidence in their favor (or any evidence at all). This generates just the sort of bias that cannot be tolerated by arguments that depend on direct inference, as the following example shows.

If the sequence Ti is a sequence of exchangeable trials (in fact all we need stipulate is that the probability of a success followed by a failure is

the same as the probability of a failure followed by a success -- much less than full exchangeability) and the prior probability of a success is .01 and the conditional probability of a success on a second trial, given a success on the first trial is .02, then we must assign a probability of at least .9996 that no more than half the trials in the arbitrarily long run will yield successes. Or, we can calculate that the probability that less than 80% of the trials in the long run will yield successes is at least (this is very conservative -- we use only Tchebycheff's inequality) .999844 = 1-.000156.

Now let us perform 16 trials. Suppose they all yield successes. Neyman has taught us that there is frequency information bearing on hypotheses about the long-run frequency of success. Specifically, whatever the actual frequency of success may be, at least 90% of the performances of 16 trials will yield results falling in what Neyman refers to as the confidence belt,[7] and the bounds of the confidence belt in this case are .80 and 1.00.

Neyman would say that we can be 90% confident that the long run success rate is in the interval [.80,1.00]. This is not a probability for Neyman, but it does correspond to a before trial relative frequency or probability. In fact, he writes: "If the confidence belt is constructed we may affirm that the point will [Neyman's emphasis!] lie inside the belt. This statement may be erroneous, but the probability of error is either equal to or less than 1 - e -- thus is as small as desired."[8]

But the gap between a frequency in a general class and the probability of a specific occurrence was exactly the gap that direct inference was supposed to be capable of crossing. Leave aside sophisticated philosophical doubts about the meaning of probability, and nothing could be more natural

than for the holder of a specific ticket in a thousand ticket lottery to say that the probability of his winning the first prize is 1/1000. Similarly, holding a sample comprising sixteen successes, nothing could be more natural than to say, since at least 90% of such trials yield Neyman-representative results, that the probability is at least 0.9 that our trial yielded a Neyman-representative result -- i.e., a result in the confidence belt. And it did this if and only if the success rate in the population is in the interval [.80,1.00].

Note that I haven't spelled out a principle of direct inference that gives this result and doesn't lead to difficulties. This isn't easy, though I think that after some 25 years, I've gotten close to it. But this is exactly the sort of thing that everybody took for granted when the problems of inverse inference were first raised. This is the sort of uncertain inference that seemed unproblematic. It is furthermore a kind of uncertain inference that even Bayesians seem to be rediscovering.[9]

Suppose the direct inference does go through. How does it relate to the previous result? Writing the appropriate form of Bayes' theorem, we have (where $\underline{r}$ is the long run relative frequency of success):

$$\underline{P}(\underline{r} > .8/\underline{E}(16,16)) =$$

$$\frac{\underline{P}(\underline{r} > .8)*\underline{P}(\underline{E}(16,16)/\underline{r} > .8)}{\underline{P}(\underline{r} > .8)*\underline{P}(\underline{E}(16,16/\underline{r} > .8) + \underline{P}(\underline{r} \leq .8)*\underline{P}(\underline{E}(16,16)/\underline{r} \leq .8)}$$

where $\underline{E}(16,16)$ is our evidence.

For consistency with direct inference, we require

$$\underline{P}(\underline{r} > .8/\underline{E}(16,16)) \leq 0.9, \text{ or}$$

$$\underline{P}(\underline{r} > .8)*\underline{P}(\underline{E}(16,16)/\underline{r} > .8) \geq .9*(\underline{P}(\underline{r} > .8)*\underline{P}(\underline{E}(16,16)/\underline{r} > .8) + \underline{P}(\underline{r} \leq .8)*\underline{P}(\underline{E}(16,16)/\underline{r} \leq .8))$$

Simplifying, and taking account of the fact that $\underline{P}(\underline{r} > .8) < .000156$ and $\underline{P}(\underline{r} \leq .8) \geq .999844$, we require:

$$\underline{P}(\underline{E}(16,16)/\underline{r} > .8) > 5.768* 10^4 \underline{P}(\underline{E}(16,16)/\underline{r} \leq .8).$$

That is, it must be nearly six thousand times as probable that we will observe $\underline{E}(16,16)$ given that $\underline{r}$ is greater than .8 than it is given that $\underline{r}$ is less than or equal to .8. This doesn't seem very plausible, but perhaps it could be argued that it just shows that our original intuitions about the frequency of successes were not as plausible as they seemed.

VI

When we formulate a principle of direct inference that allows for <u>imprecise</u> knowledge, even this sort of retroactive adjustment is impossible.

I assume that we want to apply direct inference even when we do not know <u>exactly</u> the relevant frequencies in our reference classes. This is obviously important pragmatically, but it raises a theoretical question. Suppose that we know the relevant frequency in a large class quite precisely, but that our knowledge concerning a subclass is rather vague. As an extreme example, we might know that the relative frequency of heads in coin tosses was very near 1/2; but what we know about the <u>next toss</u> is no more than that the frequency of heads is either 0 or 1. There is ordinarily a continuum of knowledge in between: tosses of U.S. coins, tosses of post-1900 coins, tosses of quarters, tosses performed on Thursdays ... at each step the size of the sample on which our knowledge might be based is smaller, and <u>ceteris paribus</u>, our knowledge becomes less precise.

On which of these various items of information whould we base our probabilities? I suggest that if there is no conflict between two items of knowledge -- if the interval corresponding to the larger class is a subinterval of the interval corresponding to the smaller class, it seems appropriate to take the smaller interval, based on the larger class, as legislative for probability.

For example, I know that very nearly a half of coin-tosses in general yield heads. I know much less about the frequency with which this particular 1980 quarter lands heads -- perhaps, by an inductive inference I could say that I know that between 40% and 60% of its tosses yield heads. But if you ask for the probability of heads on the next toss of this 1980 quarter, since there is no conflict between what I know of it and what I know of tosses in general, I shall take the narrow interval corresponding to my knowledge of tosses in general to give that probability.

This, in my systematic treatment of direct inference, is called "the strength rule". I shall now describe an example that shows that direct inference, if it incorporates the strength rule, is flatly inconsistent with inverse inference. The example is due to Isaac Levi[10] who draws the conclusion that the strength rule is unacceptable. I shall alter the example slightly, but I shall keep the numbers roughly the same. And I shall draw the opposite conclusion: that we should hold fast to the strength rule, and let inverse inference and the form of conditionalization that it requires go hang.

VII

Suppose we measure lengths with one of three instruments, A, B, and C. Instrument of type A give results accurate within a margin of error m

between 88.5% and 88.9% of the time; of type B between 90.5% and 90.9% of the time; and of type C between 91.5% and 91.9% of the time. To fix our ideas, suppose "within a margin of error m" amounts to having a reading within .001 of the true value being measured. We may suppose that these frequencies are reported by the manufacturer of the three types of instrument. In general, though, we know that the combined results of all three kinds of measurement are accurate within a margin of error m between 89.9% and 90.1% of the time. Put otherwise: we have a population of measurements of which between 89.9% and 90.1% are accurate; this population is partitioned into three sub-populations, A, B, and C, characterized by the error rates mentioned.

A particular measurement is made. We don't know what instrument was used. It seems natural to say that the probability of its being accurate within the margin of error m is (about) .90 -- more exactly, the interval [.899,.901] seems to capture what we know.

We also know that an instrument yielding an error rate between 88.5% (the minimum for instrument A) and 91.9% (the maximum for instrument C) was used, so one might be tempted to think that the appropriate interval was [.885,.919]. The strength principle argues against this; if we have more accurate information we should use it. We should use the most exact statistical knowledge we have for direct inference, provided that it is not in conflict with other knowledge that we have.

But this position is in flat-out contradiction to inverse inference construed as conditionalization on a prior probability -- i.e., as inverse inference proper. To see this, suppose that B is a Bayesian belief function, and that direct inference, as I have described it holds. Following Levi, we show that this leads, in combination with other plausible assumptions, to a contradiction.

First off, if all we know is that the measurement we have made was made with an instrument manufactured by the firm in question, we should accept the general frequency of error as constraining our epistemic probability Prob:

(*) $\underline{Prob}$ ("$\underline{S} \in \underline{G_m}$", $\underline{K_{A \cup B \cup C}}$) $\in$ [.899,.901]

where $\underline{S}$ is the particular measurement at issue, $\underline{G_m}$ is the set of measurements accurate within a margin of error $\underline{m}$, and $\underline{K_{A \cup B \cup C}}$ is the body of knowledge embodying merely the information that S was made with one of the types of instrument $\underline{A}$, $\underline{B}$, or $\underline{C}$.

Clearly, if we know which subset of the population of measurements we are in, the error rate in that subset is indicated as the appropriate basis for a direct inference.

Let $\underline{A}$, $\underline{B}$, $\underline{C}$ be the sets of all trials, past, present, and future, with instruments of types $\underline{A}$, $\underline{B}$, and $\underline{C}$, respectively.

We are warranted in accepting

(1) $\%(\underline{A},\underline{G_m}) \in [.885,.889]$

(2) $\%(\underline{B},\underline{G_m}) \in [.905,.909]$

(3) $\%(\underline{C},\underline{G_m}) \in [.915,.919]$

Let $\underline{K_A} = \underline{K} \cup \{\text{"}\underline{S} \in \underline{A}\text{"}\}$; $\underline{K_B}$, $\underline{K_C}$ similarly;

$\underline{K_{A \cup B}} = \underline{K} \cup \{\text{"}\underline{S} \in \underline{A \cup B}\text{"}\}$; $\underline{K_{B \cup C}}$, $\underline{K_{A \cup C}}$ similarly.

From (1), (2), and (3) it follows that

(4) $\%(\underline{A} \cup \underline{B},\underline{G_m}) \in [.885,.909]$

(5) $\%(\underline{A} \cup \underline{C},\underline{G_m}) \in [.885,.919]$

(6) $\%(\underline{B} \cup \underline{C},\underline{G_m}) \in [.905,.919]$

It may be the case that we have more precise knowledge of these disjunctive reference sets, as we do of $\underline{A} \cup \underline{B} \cup \underline{C}$ -- but that need not be the case. We may have lost the data; we may (reasonably) be depending on what the

manufacturer tells us; in general we cannot suppose that we know everything that anybody else knows or ever knew.

For the same reason we have, by direct inference, using the strength rule,

$$\underline{\text{Prob}}\ (\text{"}\underline{S}\varepsilon\underline{G_m}\text{"},\underline{KA}) \quad \varepsilon \quad [.885,.889]$$

$$\underline{\text{Prob}}\ (\text{"}\underline{S}\varepsilon\underline{G_m}\text{"},\underline{KB}) \quad \varepsilon \quad [.905,.909]$$

$$\underline{\text{Prob}}\ (\text{"}\underline{S}\varepsilon\underline{G_m}\text{"},\underline{KC}) \quad \varepsilon \quad [.915,.919]$$

$$\underline{\text{Prob}}\ (\text{"}\underline{S}\varepsilon\underline{G_m}\text{"},\underline{K_{A\cup B}}) \quad \varepsilon \quad [.899,.901]$$

$$\underline{\text{Prob}}\ (\text{"}\underline{S}\varepsilon\underline{G_m}\text{"},\underline{K_{A\cup C}}) \quad \varepsilon \quad [.899,.901]$$

$$\underline{\text{Prob}}\ (\text{"}\underline{S}\varepsilon\underline{G_m}\text{"},\underline{K_{B\cup C}}) \quad \varepsilon \quad [.905,.919]$$

There is no function $\underline{B}$ that's a conditional belief function such that in general

$$\underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_X}) \ \varepsilon\ \underline{\text{Prob}}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"},\underline{K_X}).$$

To see this, suppose that $\underline{B}$ is such a belief function. By conditionalization and "total probability" we have

(7) $\underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K}) = \underline{B}(\text{"}\underline{S}\varepsilon\underline{A\cup B}\text{"}/\underline{K}) * \underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_{A\cup B}}) + \underline{B}(\text{"}\underline{S}\varepsilon\underline{C}\text{"}/\underline{K}) * \underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}/\underline{K_C})\ \varepsilon$ [.899,.901].

Similarly,

(8) $\underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K}) = \underline{B}(\underline{S}\varepsilon\underline{A\cup C}\text{"}/\underline{K}) * \underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_{A\cup C}}) + \underline{B}(\underline{S}\varepsilon\underline{B}\text{"}/\underline{K})) * \underline{B}(\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_B})\ \varepsilon$ [.899,.901].

Let $\alpha = \underline{B}(\text{"}\underline{S}\varepsilon\underline{A}'/\underline{K})$; $\beta = \underline{B}(\text{"}\underline{S}\varepsilon\underline{B}\text{"}/\underline{K})$; $\gamma = \underline{B}(\text{"}\underline{S}\varepsilon\underline{C}\text{"}/\underline{K})$.

From (7) and (8), together with the principle that beliefs should be constrained by probabilities, we obtain:

$$(1-\beta)\underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_{A\cup C}}) + \beta\ \underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_B})\ \varepsilon\ [.899,.901]$$

$$(1-\gamma)\underline{B}\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_{A\cup B}}) + \gamma\ \underline{B}(\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_C})\ \varepsilon\ [.899,.901]$$

from which it follows that

$$\gamma \leq \frac{2}{17} = .1176$$

$$\beta \leq \frac{2}{7} = .2857$$

$$\alpha \geq .5967$$

Note that α, β, and γ are <u>not</u> probabilities, based on known frequencies, but mere degrees of belief, based on the principle of direct inference <u>together with</u> the probability calculus.

Given these constraints on α, β, and γ, we may derive a constraint on $\underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K})$.

$$\underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K}) = \underline{B}(\text{"}\underline{S}\varepsilon\underline{A}\text{"}/\underline{K}) \cdot \underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_A}) +$$
$$\underline{B}(\text{"}\underline{S}\varepsilon\underline{B}\text{"}/\underline{K}) \cdot \underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}/\underline{K_B}) +$$
$$\underline{B}(\text{"}\underline{S}\varepsilon\underline{C}/\underline{K}) \cdot \underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_C}).$$

The <u>maximum</u> possible value for $\underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K})$ given these constraints is

$$\alpha_{\underline{min}} \cdot \underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_A}) + \beta_{\underline{max}}\, \underline{B}(\text{"}\underline{S}\varepsilon\underline{G_m}\text{"}\underline{K_B})$$
$$+ \gamma_{\underline{max}} \cdot \underline{B}(\underline{S}\varepsilon\underline{G_m}\text{"}/\underline{K_C}) \leq$$
$$.5967(.889) + .2857(.909) + .1176(.919) =$$
$$.5305 + .2597 + .1801 = .8982.$$

But this does not fall within the constraints imposed by the principle of direct inference, <u>viz</u>., [.899,.901].

It might be suggested that these are just not plausible statistics for us to know. The response is, first, that these may just be the statistics we have to work with. The second response is that even if we must get the statistics from our own data, we can generate the problem.

Pick a level of acceptance -- e.g., $1 - \alpha = 0.99$. Look up a number <u>n</u> such that a .99 confidence interval based on a sample of size <u>n</u>, with observed relative frequency of $\underline{G_m}$ about 0.917, is included in [915,.919].

Since the .99 confidence interval corresponds to about 2.575 standard deviations (using a normal approximation and $1/2\sqrt{n}$ as the upper bound of the standard deviation, n is about 414414. Similarly, for A and for B, about the same n will do. To obtain an overall confidence interval of [.899,.901], we may suppose a further, undifferentiated sample of 413395, of which 367498 are G_m. (Not all possible data is recorded; not all recorded data is kept.)

Even if we try to be realistic about the data, we encounter the conflict. But we have no reason to suppose we have the data -- the error rates may just be reported in this form by the manufacturer.

VII

This just exhibits one more conflict between direct inference and inverse inference. What do we do about it? One answer is to circumscribe direct inference enough so that it can be reconciled with inverse inference. One way to do this is to obtain probabilities from statistical knowledge only when they concern objects (or events or whatever) that are random members of their appropriate reference classes. This is Levi's suggestion.[11] But to construe randomness in this way is, as I see it, to abandon direct inference. We do not obtain the probability of accuracy of our measurement from knowledge of the frequencies of error in $A \cup B \cup C$ and its subsets A, B, and C, but from that statistical knowledge combined with non-statistical "probabilistic" knowledge about how the measurement was generated. (Clearly if knowledge about how the measurement was generated is statistical, we face no problem; but then all probabilities can come from direct inference.)

The most important philosophical counterargument to sacrificing conditionalization to direct inference is the Dutch Book argument. Just as

it is alleged that one's degrees of belief should satisfy the axioms of the probability calculus, else one could have a book made against one, so, it is argued, if conditional bets are allowed, one's conditional degrees of belief must satisfy the principle of conditionalization. More explicitly, suppose that the interval of probability for S is [.3,.4] and that for S & T is [.1,.2]. Then the Bayesian conditional probability of T given S should be constrained by the interval [.26,.67]. Every classical probability function P such that P(S) ε [.3,.4] and P(S & T) ε [.1,.2] is such that P(T/S) ε [.25,.67].

Clearly the interval [.25,.67] should constrain the odds of conditional bets on T given S. It is claimed that the same interval should constrain my bets on T after I have added S to my body of knowledge. This principle is one that Levi has called "confirmational conditionalization". As was first pointed out by him, and as we have just seen, confirmational conditionalization is in conflict with at least some forms of direct inference.

Suppose we abandon confirmational conditionalization, as I have suggested. Then after observing S the (new) probability of T need not be the interval [.25,.67] -- or any subinterval of it -- but might be (say) [.70,.80]. The cunning bettor, knowing that I will modify my probabilities in this way, offers a bet at odds of 4 to 6 against S, and also a conditional bet at even money on T given S for a stake of 11. Then, knowing how I will modify my odds on learning S, he plans to make a new bet after S has occured (if it does) against T, at 6 units to 18. Here is what happens:

If S fails to occur, the bettor gains 6 units and no other bets are activated.

If S does occur, there are two cases. If T occurs the bettor loses 4 from the first bet; gains 11 from the conditional bet and loses 6 from the

third bet, for a net gain of 1. If T fails to occur, he loses 4 from the first bet; loses 11 from the conditional bet, but gains 18 from the third bet for a net gain of 3. In any case, the bettor wins. I have been Dutch Booked!

By giving up confirmational conditionalization, have I not laid myself open to a sure loss? Of course not. Even in the elementary case, I could be willing to offer odds of 2 to 1 on S and 2 to 1 against S without being willing to make both bets at once. But even if I must post odds, and must take any bets consistent with that posting, so that the posted odds must be coherent, on pain of sure loss, that is no argument that I cannot change my posting. (In fact, the willingness to change the posted odds in the face of new evidence might be one of the things that distinguishes successful bookies!)

But we must be careful about the sort of changes that evidence can warrant. The case at hand is not one that can actually happen. The clue to this lies in the fact that the probabilities mentioned entail that P(S & -T) = 0.2 exactly; the probability that I would assign to T after observing S can be construed as a constraint on the prior probabilities of S & T and S & -T. It is only where the strength rule is invoked that we can have a legitimate violation of confirmational conditionalization. It is easy to see this in the original example concerning the three measuring instruments A, B, and C. If we hang on to the probabilities that are determined by direct inference, there is NO adjustment of prior probabilities that will preserve coherence. To regain coherence, at least one direct inference must go!

The ideal Bayesian robot, to be sure, has no need of either direct inference or interval-valued probabilities. But his probabilities are at base *a priori* prejudices, compounded, to be sure, with observations.

Alternatively, we may follow Fisher and Neyman in abandoning inverse inference. Especially once we have liberalized our notion of probability to accommodate intervals or sets of distributions, the loss of inverse inference is no loss.

The basic Bayesian Blunder does not lie in the use of Bayes' theorem. The use of Bayes' theorem is perfectly compatible with the principle that all probabilities, without exception, are obtained by direct inference. The blunder lies in the conviction that only by inverse inference proper can the knowledge needed for direct inference be obtained. But we can't get acceptance from inverse inference alone, so inverse probability doesn't solve that problem. And, worse, inverse inference is seriously incompatible with direct inference, which was where we started from. The whole idea of inverse inference was to complete and complement an acceptable theory of direct inference; what we find, when we develop inverse inference far enough is that we have little or nothing left of direct inference. We have undermined the foundation on which we tried to build.

Henry E. Kyburg, Jr.
University of Rochester

FOOTNOTES

* Research for this paper was supported in part by the U.S. Army Signals Warfare Laboratory.

1. "Uncertain Inference," Proceedings of the American Academy of Arts and Sciences 71, 1936, pp. 245-258, p.254.

2. Neyman, Jerzy, "'Inductive Behavior' as a Basic Concept as of Philosophy of Science," Bulletin of the International Statistical Institute 25, 1957, Pp. 7-22. "Raisonnement Inductif ou Comportement Inductif, Proceedings of the International Statistical Conference 1947, 3, 423-433.

3. "Fiducial Argument and the Theory of Confidence Intervals," Biometrika 32, 1941, pp. 128-150, p. 378.

4. "Inverse Probability," Proceedings of the Cambridge Philosophical Society 26, 1930, pp. 528-535.

5. "Inverse Probability," Proceedings of the Cambridge Philosophical Society 26, 1930, pp. 528-535.

6. Many writers insist that one can make probability statements only about random quantities and not about constants. But, for example, $(\underline{x} - \mu)/\sigma$ is a perfectly good random quantity, and yet to say that its absolute value is less than $\underline{e}$ is exactly to say that $\underline{x} - \sigma\,\underline{e} < \mu < \underline{x} + \sigma\,\underline{e}$, which is a statement "about" the constant μ. Of course there are epistemological problems to be dealt with before we can go around substituting observed values for the quantity $\underline{x}$, but that's another matter.

7. Neyman, "On the Two Different Aspects of the Representative Method," Journal of the Royal Statistical Soceity 97, 1934, pp. 558-625.

8. Op. cit. p. 123.

9. For example, David Lewis, "A Subjectivists' Guide to Objective Chance," in R.C. Jeffrey (ed.), Studies in Inductive Logic and Probability, University of California Press, 1980, pp. 263-294, proposes a "principal principle" that is none other than a newly discovered principle of direct inference.

10. "Direct Inference," Journal of Philosophy, 74, 1977, pp. 5-29, also The Enterprise of Knowledge. MIT Press, Cambridge, 1980.

11. Ibid.